A New Confidence Interval for the Odds Ratio: an Application to the Analysis of the Risk of Survival of an Enterprise

Wojciech Zieliński[1]

Abstract

We consider the problem of interval estimation of the odds ratio. An asymptotic confidence interval is widely applied. Unfortunately that confidence interval has a poor coverage probability: it is significantly smaller than the nominal confidence level. In this paper a new confidence interval is proposed. The coverage probability of the proposed confidence interval is at least the nominal confidence level.

The new confidence interval is applied to the analysis of the risk of liquidation of a company during the first ten years of its activity. Companies established in 2007 in Mazowieckie voivodship and Warsaw are analysed with respect to their surviving during the first ten years of activity.

Keywords: odds ratio, confidence interval, survival of firms

JEL Classification: C13, C19

1. Introduction

In practical applications we often need to compare two groups using binary data. Three parameters are commonly used: the difference of two proportions (the risk difference), the ratio of two proportions (the relative risk), and the odds ratio. The odds ratio is one of the parameters commonly used in such comparisons. This indicator was firstly applied by Cornfield (1951). The literature devoted to the analysis of odds ratio and its estimators is very rich, see e.g. Encyclopedia of Statistical Sciences and the literature therein.

However, the problem is in the interval estimation. There are two approaches to the problem. The first one consists of the analysis of 2×2 tables (Edwards, 1963; Gart, 1971; Thomas, 1971). The second approach is based on logistic model in which the odds ratio has a direct relationship with the regression coefficient (Gart, 1971; McCullagh, 1980; Morris and Gardner, 1988). This approach is commonly used in applications and an asymptotic interval for odds ratio derived from the logistic model is widely used. This interval is applied in different statistical packages. There are also many internet scripts for calculating an asymptotic confidence interval (see e.g. http://www.hutchon.net/ConfidOR.htm). Unfortunately this

[1] Corresponding author: Warsaw University of Life Sciences, Department of Econometrics and Statistics, Nowoursynowska 159, PL-02-776 Warsaw, wojciech_zielinski@sggw.pl, http://wojtek.zielinski.statystyka.info.

confidence interval has some statistical disadvantages discussed (Zieliński, 2019). To avoid those disadvantages a new confidence interval is proposed. The new confidence interval was applied to the analysis of surviving of firms.

In the analysis of surviving of firms an important indicator is the probability of their liquidation. Such probability is estimated with the aid of the Kaplan-Meier survival estimator in different groups of risks. An exhaustive analysis may be found in Łobos and Szewczyk (2013), Markowicz (2018, 2019), Ptak-Chmielewska (2013), Ptak-Chmielewska and Matuszyk (2019).

In what follows, we are interested only in the comparison of the risk of surviving of firms established in Mazowieckie voivodship and Warsaw. All investigated firms were established in 2007 and it was observed whether they were still active in 2017 (i.e. after ten years).

2. Confidence interval for odds ratio

Consider two independent r.v.'s ξ_A and ξ_B distributed as Bin(n_A, p_A) and Bin(n_B, p_B), respectively. The problem is in estimating the odds ratio:

$$OR = \left(\frac{p_A}{1-p_A}\right)\left(\frac{1-p_B}{p_B}\right).$$

Let n_{A1} and n_{B1} be observed numbers of successes. The data are usually organized in a 2×2 table:

Table 1. 2×2 table of data

	Success	**Failure**	
A	n_{A1}	n_{A0}	n_A
B	n_{B1}	n_{B0}	n_B
	n_1	n_0	n

The standard estimator of OR is as follows:

$$\widehat{OR} = \left(\frac{n_{A1}}{n_A - n_{A1}}\right)\left(\frac{n_B - n_{B1}}{n_{B1}}\right). \quad (1)$$

The standard asymptotic confidence interval at the confidence level γ for odds ratio is of the form

$$\left(\widehat{OR}\, Exp\left(u_{\frac{1-\gamma}{2}}S\right), \widehat{OR}\, Exp\left(u_{\frac{1+\gamma}{2}}S\right)\right) \quad (2)$$

where:

$$S = \sqrt{\frac{1}{n_{A1}} + \frac{1}{n_A - n_{A1}} + \frac{1}{n_{B1}} + \frac{1}{n_B - n_{B1}}}$$

Here u_γ is the γ quantile of the $N(0,1)$ distribution.

This confidence interval has at least two disadvantages. It does not exist when

$$n_{A1} = 0 \; or \; n_A - n_{A1} = 0 \;\; or \; n_{B1} = 0 \;\; or \; n_B - n_{B1} = 0.$$

Also, its coverage probability is less than the nominal one (Zieliński, 2019).

Note that the estimator $\widehat{OR}$ given by the formula (1) is undefined for $n_{A1} = 0 \; or \; n_{A1} = n_A \; and \; n_{B1} = 0 \; or \; n_{B1} = n_B$. We extend the definition of OR in the following way:

$$\widehat{OR} = \begin{cases} 0, (\mathrm{n_{A1}} = 0, \mathrm{n_{B1}} \geq 1) \text{ or } (\mathrm{n_{A1}} \geq 1, \mathrm{n\_(B1)} = \mathrm{n_B}) \\ +\infty, (\mathrm{n_{A1}} = \mathrm{n_A}, \mathrm{n_{B1}} \geq 1) \text{or } (\mathrm{n_{A1}} \leq \mathrm{n_A} - 1, \mathrm{n_{B1}} = 0) \\ 1, (\mathrm{n_{A1}} = 0, \mathrm{n_{B1}} = 0) \text{or } (\mathrm{n_{A1}} = \mathrm{n_A}, \mathrm{n_{B1}} = \mathrm{n_B}) \\ \text{formula } (*), \text{elsewhere} \end{cases}$$

To find the distribution of $\widehat{OR}$ note that for a given odds ratio equal to $r > 0$ we have $p_B = \frac{p_A}{p_A + r(1-p_A)}$.

The probability of observing $\xi_A = n_{A1}$ and $\xi_B = n_{B1}$ equals

$$P_{r,p_A}\{n_{A1}, n_{B1}\} = r^{n_B - n_{B1}} \binom{n_A}{n_{A1}} \binom{n_B}{n_{B1}} \frac{p_A^{n_{A1}+n_{B1}} (1-p_A)^{n_A + n_B - n_{A1} - n_{B1}}}{\left(p_A + r(1-p_A)\right)^{n_B}}.$$

The probability p_A is eliminated by an appropriate integration:

$$P_r\{n_{A1}, n_{B1}\} = \int_0^1 P_{r,p_A}\{n_{A1}, n_{B1}\} dp_A.$$

The cumulative distribution function of $\widehat{OR}$ has the form

$$F_r(t) = P_r\{\widehat{OR} \leq t\} = \sum_{n_{A1}=0}^{n_A} \sum_{n_{B1}=0}^{n_B} \mathrm{P_r}\{\mathrm{n_{A1}}, \mathrm{n_{B1}}\} \, \mathbf{1}\left(\widehat{OR}(n_{A1}, n_{B1}) \leq t\right)$$

where $\mathbf{1}(q) = 1$ when q is true and $= 0$ elsewhere.

Let $G_r(t) \; = P_r(\widehat{OR} < \; t)$.

Let γ be the given confidence level and let $\hat{r}$ be observed odds ratio. The confidence interval for r takes on the form

$$(Left(\hat{r}), Right(\hat{r})), \tag{3}$$

where

$$Left(\hat{r}) = \begin{cases} 0, & \hat{r} = 0, \\ 0, if \lim\limits_{r\to 0} G_r(\hat{r}) < \dfrac{1+\gamma}{2} \\ r_*, r_* = \max\{r: G_r(\hat{r})\} \geq \dfrac{1+\gamma}{2} \end{cases}$$

It is interesting to note, that for $n_A > \frac{2}{1-\gamma} - 1$ the confidence interval is two-sided, and it is one-sided otherwise.

Unfortunately, no closed formulae for the ends of the confidence interval are available. However, for given n_A, n_B and observed $\widehat{OR}$ the ends of the confidence interval may be easily numerically computed with the aid of the standard software such as R, Mathematica etc. The proposed confidence interval may be applied for small as well as large sample sizes.

3. Application

The aim of the study was to compare the chances of survival of trading companies in Mazowieckie voivodship versus Warsaw. The question was about the chances of surviving during the first ten years of activity.

Let p_A denote the probability of surviving the first ten years of activity of a firm established in Mazowieckie voivodoship, and let p_B denote the appropriate probability for a firm established in Warsaw. We are interested in estimation of the odds ratio, i.e. $\frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}$.

From the REGON register it is known that 32760 firms started their activity in 2007. Among them 17130 were established in Mazowieckie voivodship, while 15630 were established in Warsaw.

Among firms established in 2007 the random sample of size 320 was taken and it was observed how many of those firms were still active in 2017. The data are given in Table 2.

Table 2. Random sample of firms

	Active	**Nonactive**	
Mazowieckie	96	74	170
Warsaw	85	65	150

On the basis of those data the odds ratio would be estimated.

Note that the estimator of the odds ratio is defined for random variables distributed as binomial. In our investigation we deal with random variables distributed as hypergeometric. It is well known that hypergeometric distribution may be approximated by an appropriate binomial distribution. Some remarks on consequences of such approximation may be found in Zieliński (2011). In what follows, it is assumed that binomial approximation to the hypergeometric one is fairly enough.

The estimate of odds for Mazowieckie voivodship equals

$$\frac{\frac{96}{170}}{\frac{74}{170}} = 1.297.$$

It means that almost 30% more of the firms established in 2007 were still working than were nonactive. Similar indicator for Warsaw equals 1.308.

The estimate of odds ratio for Mazowieckie voivodship versus Warsaw equals

$$\frac{1.292}{1.308} = 0.992.$$

The confidence interval (3) at 95% confidence level is (0.437, 2.049). Since this confidence interval covers 1, it may be expected that for the firms established in 2007 the chances of surviving the first ten years of activity for Mazowieckie voivodship and for Warsaw are similar. The above conclusion may, of course, be wrong. It must be stressed that the risk of over- or under-estimation is at most 5%, in contradiction to the standard confidence interval.

Simple calculations show that the standard confidence interval (2) at 95% confidence level for odds ratio is (0.989, 1.544). This confidence interval is narrower than (3), but unfortunately the risk of not covering the true value of the odds ratio is greater than assumed 5% and remains unknown (Zieliński, 2019).

In the presented example we are very lucky since we have full information about the number of firms established in 2007 which survived till 2017. Hence, we may calculate the exact value of odds ratio for that population. Those data are presented in Table 3 (data comes from the REGON register)

Table 3. Exact numbers of firms

	Active	**Nonactive**	
Mazowieckie	9448	7682	17130
Warsaw	9607	6023	15630

The exact value of odds ratio in that population equals:

$$\frac{9448/7682}{9607/6023} = 0.771.$$

Note that the new confidence interval (3) covers this value, while the standard asymptotic confidence interval does not.

4. Conclusions

A new confidence interval for odds ratio is proposed. This confidence interval may be applied for small as well as large samples. By construction, the probability of coverage of the true value of odds ratio is at least equal to nominal confidence level. The standard and widely applied confidence interval has the coverage probability less than the nominal confidence level. As a consequence the risk of wrong conclusion resulting in applying the standard confidence interval is significantly greater than assumed.

Acknowledgements

All data are presented with the kind permission of Mrs Dominika Urbańczyk (Department of Econometrics and Statistics, Warsaw University of Life Sciences).

References

Cornfied, J. (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix. *JNCI: Journal of the National Cancer Institute*. 11:1269-1275, DOI: 10.1093/jnci/11.6.1269.

Edwards, A.W.F. (1963). The Measure of Association in a 2×2 Table. *Journal of the Royal Statistical Society. Ser. A.* 126: 109–114. DOI: 10.2307/2982448.

Gart, J.J. (1971). The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification. *Review of the International Statistical Institute*. 39: 148-169.

Łobos, K., Szewczyk, M. (2013). Survival analysis: a case study of micro and small enterprises in Dolnoslaskie and Opolskie Voivodship, *Central and Eastern European Journal of Management and Economics*, 1, 123-140.

Encyclopedia of Statistical Sciences (http://www.mrw.interscience.wiley.com/ess), Volume 9, pp. 5722-5726.

Markowicz, I. (2018). Modeling the Survival Time of Trading *Companies in the Zachodniopomorskie Voivodship*, 4(337), 85-97, DOI: 10.18778/0208-6018.337.06.

Markowicz, I. (2019). Analysis of the risk of liquidation depending on the age of the company: a study of entities established in Szczecin in period 1990-2010. *EKONOMETRIA*, 23, 49-62, DOI: 10.15611/eada.2019.2.04.

Morris, J.A., Gardner, M.J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal*. 296: 1313-6. DOI: 10.1136/bmj.296.6632.1313.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Ser. B.* 42: 109-142.

Ptak-Chmielewska, A. (2013). Semiparametric Cox regression model in estimation of small and micro enterprises' survival in the Malopolska voivodeship. *Metody Ilościowe w Badaniach Ekonomicznych*, XIV, 163-180.

Ptak-Chmielewska, A., Matuszyk, A. (2019). Macroeconomic factors in modelling the smes bankruptcy risk. The case of the polish market. *EKONOMETRIA*, 23, 40-49, DOI: 10.15611/eada.2019.3.04.

Thomas, D.G. (1971). Algorithm AS-36: exact confidence limits for the odds ratio in a 2×2 table. *Applied Statistics*. 20: 105-110.

Zieliński, W. (2011). Comparison of confidence intervals for fraction in finite populations. *Metody Ilościowe w Badaniach Ekonomicznych*, XII, 177-182.

Zieliński, W. (2019). A New Confidence Interval for the Odds Ratio, http://arxiv.org/abs/1910.03832.